Supplementary materials for

Heavy Ball Momentum Does Not Always Accelerate SGD

.1. Proof of Lemma 3.1

We need to exploit the eigenvalues of $\mathcal{T}$, i.e., the complex number λ satisfying

$$\det\begin{pmatrix} (\lambda-1-\beta)\mathbf{I}+\gamma\mathbf{A} & \beta\mathbf{I} \\ -\mathbf{I} & \lambda\mathbf{I}\end{pmatrix}=0.$$

Then we have

$$\det\begin{pmatrix} (\lambda+\frac{\beta}{\lambda}-1-\beta)\mathbf{I}+\gamma\mathbf{A} & \mathbf{0} \\ -\mathbf{I} & \lambda\mathbf{I}\end{pmatrix}=0$$
$$\Longrightarrow \det((\lambda+\frac{\beta}{\lambda})\mathbf{I}-[(1+\beta)\mathbf{I}-\gamma\mathbf{A}])=0.$$

If λ^* is a eigenvalue of $\mathbf{A}$, we just need to consider

$$\lambda+\frac{\beta}{\lambda}=(1+\beta)\mathbf{I}-\gamma\lambda^*. \tag{8}$$

Let $\mathbf{U}:=[\mathbf{u}_1,\mathbf{u}_2,\dots,\mathbf{u}_d]$ be the eigenvectors of $\mathbf{A}$, it then holds

$$\langle\mathbf{u}_i,\mathbf{u}_j\rangle=0 \text{ if } i\neq j,$$

since $\mathbf{A}$ is symmetry positive definite. It is easy to see that $\mathbf{u}_1,\mathbf{u}_2,\dots,\mathbf{u}_d$ are the eigenvectors of $(1+\beta)\mathbf{I}-\gamma\mathbf{A}$. Let λ_i be the ith eigenvalue of $\mathbf{A}$, we can see

$$(1+\beta-\gamma\lambda_i)^2-4\beta\leq(1+\beta-\gamma\nu)^2-4\beta\leq 0.$$

Thus, we define $\overline{\lambda_i}$ and $\underline{\lambda_i}$ as follows

$$\overline{\lambda_i}:=\frac{(1+\beta-\gamma\lambda_i)+\sqrt{4\beta-(1+\beta-\gamma\lambda_i)^2}\mathbf{i}}{2},$$

$$\underline{\lambda_i}:=\frac{(1+\beta-\gamma\lambda_i)-\sqrt{4\beta-(1+\beta-\gamma\lambda_i)^2}\mathbf{i}}{2},$$

where $\mathbf{i}^2=-1$. Direct calculating gives us

$$\mathcal{T}\begin{pmatrix}\overline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix}=\overline{\lambda_i}\begin{pmatrix}\overline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix},\mathcal{T}\begin{pmatrix}\underline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix}=\underline{\lambda_i}\begin{pmatrix}\underline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix}.$$

Therefore, all the eigenvectors of $\mathcal{T}$ can be written as

$$\left\{\begin{pmatrix}\overline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix},\begin{pmatrix}\underline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix}\right\}_{1\leq i\leq d}.$$

If $i\neq j$, we have

$$\left\langle\begin{pmatrix}\overline{\lambda_i}\mathbf{u}_i\\ \mathbf{u}_i\end{pmatrix},\begin{pmatrix}\underline{\lambda_i}\mathbf{u}_j\\ \mathbf{u}_j\end{pmatrix}\right\rangle=0.$$

Note that $\mathcal{T}$ has $2d$ different eigenvectors since $\beta=[1-\sqrt{\gamma\nu}]^2+\varrho$, $\overline{\lambda_i}\neq\underline{\lambda_i}$. Denote that

$$\overline{\Lambda}:=\mathrm{Diag}(\overline{\lambda_1},\overline{\lambda_2},\dots,\overline{\lambda_d}),\ \underline{\Lambda}:=\mathrm{Diag}(\underline{\lambda_1},\underline{\lambda_2},\dots,\underline{\lambda_d}).$$

We then construct the following matrix

$$\mathcal{U}:=\begin{pmatrix}\overline{\Lambda}\mathbf{U} & \underline{\Lambda}\mathbf{U}\\ \mathbf{U} & \mathbf{U}\end{pmatrix}.$$

It is known that $\mathcal{T}$ can be decomposed as

$$\mathcal{T} = \mathcal{U}^{-1} \begin{bmatrix} \overline{\Lambda} & \\ & \underline{\Lambda} \end{bmatrix} \mathcal{U}. \tag{9}$$

Therefore,

$$\mathcal{T}^k = \mathcal{U}^{-1} \underbrace{\begin{bmatrix} \overline{\Lambda}^k & \\ & \underline{\Lambda}^k \end{bmatrix}}_{:=\Lambda} \mathcal{U}.$$

We are then led to

$$\|\mathcal{T}^k\| = \|\mathcal{U}\Lambda^k\mathcal{U}^{-1}\| \le \|\mathcal{U}\|_F\|\mathcal{U}^{-1}\|_F \cdot 2d|\lambda_{\max}|^k, \tag{10}$$

where we use the fact that $\|\mathbf{M}\mathbf{N}\|_F \le \max\{\|\mathbf{M}\|_F\|\mathbf{N}\|, \|\mathbf{M}\|\|\mathbf{N}\|_F\}$. When $\beta = (1-\sqrt{\gamma\nu})^2 + \varrho$ and $0 < \varrho \ll \epsilon$,

$$|\lambda_{\max}| \le 1 - \sqrt{\gamma\nu} + \varrho.$$

Direct calculation yields

$$\mathcal{U}^{-1} := \begin{pmatrix} \mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1} & -\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\underline{\Lambda} \\ -\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1} & \mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\overline{\Lambda} \end{pmatrix}.$$

From the form of $\overline{\Lambda}, \underline{\Lambda}$, we have

$$[(\overline{\Lambda}-\underline{\Lambda})^{-1}]_{i,i} = ([\overline{\Lambda}-\underline{\Lambda}]_{i,i})^{-1} = -\frac{1}{\sqrt{4\beta-(1+\beta-\gamma\lambda_i)^2}}\mathbf{i} \approx -\frac{1}{\sqrt{\gamma\nu}}\mathbf{i}.$$

That means

$$(\overline{\Lambda}-\underline{\Lambda})^{-1} \approx \frac{-\mathbf{i}}{\sqrt{\gamma\nu}}\mathbf{I}.$$

Noticing that β is very closed to 1 and ϵ is very small,

$$\underline{\Lambda} \approx \mathbf{I},\ \overline{\Lambda} \approx \mathbf{I}.$$

Turning back to $\mathcal{U}$ and $\mathcal{U}^{-1}$, we see that $\|\mathcal{U}\|_F = \mathcal{O}(1)$ and $\|\mathcal{U}^{-1}\|_F = \Theta(\frac{1}{\sqrt{\gamma\nu}})$.

.2. Proof of Lemma 3.2

If $\beta = 1-\Theta(\gamma^\tau)$ and $\tau \ge 1$, we have $\beta \ge (1-\sqrt{\gamma\nu})^2$ when γ is small, it holds $(1+\beta)-\gamma\nu \le 2\sqrt{\beta}$, the equation (8) has complex roots whose norms are both β. Thus

$$|\lambda_i| = \sqrt{\beta} \ge 1-\Theta(\gamma^\tau),\ 1 \le i \le 2d.$$

With such a choice, we still have $(\overline{\Lambda} \approx \underline{\Lambda})^{-1}$, and $\|\mathcal{U}\|_F = \Theta(1)$. Here, the norm $\|\cdot\|_F$ and $\|\cdot\|$ are taken on the complex domain. Let $\xi = (\xi_1 \in \mathbb{R}^d, \mathbf{0})$ and $\xi_1 \sim \mathcal{E}$. Denote $\bar{\xi} := \mathcal{U}\xi = \begin{bmatrix} \overline{\Lambda}\mathbf{U}\xi_1 \\ \mathbf{U}\xi_1 \end{bmatrix}$. We then have

$$\begin{aligned} \mathbb{E}\|\mathcal{T}^k\xi\|^2 &= \mathbb{E}\|\mathcal{U}^{-1}\Lambda\bar{\xi}\|^2 \ge \mathbb{E}\|\Lambda\bar{\xi}\|^2/\|\mathcal{U}\|_F^2 = \mathbb{E}\left\|\begin{bmatrix} \overline{\Lambda}^{k+1}\mathbf{U}\xi_1 \\ \underline{\Lambda}^k\mathbf{U}\xi_1 \end{bmatrix}\right\|^2/\|\mathcal{U}\|_F^2 \\ &\ge \mathbb{E}\|\underline{\Lambda}^k\mathbf{U}\xi_1\|^2/\|\mathcal{U}\|_F^2 \ge [1-\Theta(\gamma^\tau)]^{2k}\mathbb{E}\|\mathbf{U}\xi_1\|^2/\|\mathcal{U}\|_F^2, \end{aligned}$$

Since $\mathrm{Tr}(\mathbf{U}\Sigma\mathbf{U}^\top) = \mathrm{Tr}(\mathbf{U}^\top\mathbf{U}\Sigma) = \mathrm{Tr}(\Sigma)$, we then get

$$\mathbb{E}\|\mathbf{U}\xi_1\|^2 = \mathrm{Tr}(\Sigma).$$

Therefore, we have

$$\mathbb{E}\|\mathcal{T}^k\xi\|^2 \ge \mathrm{Tr}(\Sigma)/\|\mathcal{U}\|_F^2[1-\Theta(\gamma^\tau)]^{2k} = \Theta(1)[1-\Theta(\gamma^\tau)]^{2k}.$$

.3. Proof of Lemma 3.5

Let λ_i be the ith eigenvalue of $\mathbf{A}$ and $0 \le \beta \le \beta_0 < 1, 1-\beta_0 \gg \epsilon$, we can see

$$(1+\beta-\gamma\lambda_i)^2 - 4\beta \ge (1+\beta-\gamma L)^2 - 4\beta \ge 0.$$

Thus, we define $\overline{\lambda_i}$ and $\underline{\lambda_i}$ as follows

$$\overline{\lambda_i} := \frac{(1+\beta-\gamma\lambda_i) + \sqrt{(1+\beta-\gamma\lambda_i)^2 - 4\beta}}{2},$$

$$\underline{\lambda_i} := \frac{(1+\beta-\gamma\lambda_i) - \sqrt{(1+\beta-\gamma\lambda_i)^2 - 4\beta}}{2}.$$

Then, we have

$$[(\overline{\Lambda} - \underline{\Lambda})^{-1}]_{i,i} = ([\overline{\Lambda} - \underline{\Lambda}]_{i,i})^{-1} = -\frac{1}{\sqrt{(1+\beta-\gamma\lambda_i)^2 - 4\beta}} \approx \frac{1}{1-\beta}.$$

That means

$$\|(\overline{\Lambda} - \underline{\Lambda})^{-1}\| = \Theta(\frac{1}{1-\beta_0}),\ \|\overline{\Lambda}\| = \mathcal{O}(1),\ \|\underline{\Lambda}\| = O(1).$$

On the other hand,

$$\frac{(1+\beta-\gamma\lambda_i) + \sqrt{(1+\beta-\gamma\lambda_i)^2 - 4\beta}}{2} \le 1 - \frac{\gamma\lambda_i}{1-\beta} + C_3\epsilon^2,$$

which means

$$|\lambda_{\max}| \le 1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2.$$

.4. Proof of Theorem 3.3

Let

$$\mathbf{y}^k := \left[\begin{array}{c} \mathbf{w}^k - \mathbf{w}^* \\ \mathbf{w}^{k-1} - \mathbf{w}^* \end{array}\right] \in \mathbb{R}^{2d}.$$

According to the fact that $\nabla R_S(\mathbf{w}^*) = \mathbf{0}$, we have

$$\begin{aligned} \mathbf{w}^{k+1} - \mathbf{w}^* &= \mathbf{w}^k - \mathbf{w}^* - \gamma(\nabla R_S(\mathbf{w}^k) - \nabla R_S(\mathbf{w}^*)) + \beta(\mathbf{w}^k - \mathbf{w}^*) - \beta(\mathbf{w}^{k-1} - \mathbf{w}^*) + \gamma(\mathbf{g}^k - \nabla R_S(\mathbf{w}^k)) \\ &= \mathbf{w}^k - \mathbf{w}^* - \gamma\mathbf{A}(\mathbf{w}^k - \mathbf{w}^*) + \beta(\mathbf{w}^k - \mathbf{w}^*) - \beta(\mathbf{w}^{k-1} - \mathbf{w}^*) + \gamma(\mathbf{g}^k - \nabla R_S(\mathbf{w}^k)), \end{aligned}$$

where $\mathbf{A} := \nabla^2 R_S$. Then SHB can be reformulated as

$$\mathbf{y}^{k+1} = \mathcal{T}\mathbf{y}^k + \mathbf{e}^k,$$

where $\mathbf{e}^k := \begin{pmatrix} \gamma(\mathbf{g}^k - \nabla R_S(\mathbf{w}^k)) \\ \mathbf{0} \end{pmatrix}$. It is easy to see that $\mathbf{A}$ is symmetry positive definite due to the quadratic property of R_S. We then have

$$\mathbf{y}^{k+1} = \mathcal{T}^k\mathbf{y}^1 + \sum_{i=1}^{k} \mathcal{T}^{k-i}\mathbf{e}^i.$$

Using the fact that $\mathbb{E}\langle \mathbf{e}^i, \mathbf{e}^j\rangle = 0$ if $i \ne j$, we have

$$\mathbb{E}\|\mathbf{y}^{k+1}\|^2 = \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1 + \sum_{i=1}^{k} \mathcal{T}^{k-i}\mathbf{e}^i\|^2 = \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1\|^2 + \sum_{i=1}^{k} \|\mathcal{T}^{k-i}\mathbf{e}^i\|^2. \tag{11}$$

With Lemma 3.1, it follows

$$\mathbb{E}\|\mathbf{w}^K - \mathbf{w}^*\|^2 \le \mathbb{E}\|\mathbf{y}^K\|^2 \le \frac{C_1^2}{\gamma\nu}[1-\sqrt{\gamma\nu}]^{2K}\|\mathbf{y}^1\|^2 + \frac{\gamma C_1^2\sigma^2}{\nu}\sum_{i=1}^{K}[1-\sqrt{\gamma\nu}]^{2K-2i}.$$

When γ is small, $\sum_{i=1}^{K}[1-\sqrt{\gamma\nu}]^{2K-2i} \le \frac{1}{\sqrt{\gamma\nu}}$, we then proved the result.

.5. Proof of Theorem 3.4

Noticing that with Lemma 3.2, it holds $\mathbb{E}\|\mathcal{T}^{K-i}\mathbf{e}^i\|^2 \geq C_2 \cdot \gamma^2(1-\Theta(\gamma^\tau))^{2K-2i}$. Stating from (11), we are then led to

$$\mathbb{E}\|\mathbf{y}^K\|^2 \geq \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1\|^2 + C_2\gamma^2\sum_{i=1}^{K}[1-\Theta(\gamma^\tau)]^{2K-2i} = \Theta(\gamma^{2-\tau}).$$

The above equation indicates that

$$\mathbb{E}\|\mathbf{w}^K - \mathbf{w}^*\|^2 + \mathbb{E}\|\mathbf{w}^{K-1} - \mathbf{w}^*\|^2 \geq \Theta(\gamma^{2-\tau}). \tag{12}$$

According to (12), if we set $\gamma = \Theta(\epsilon)$, the lower bound is in the order of $\Theta(\epsilon^{2-\tau})$.

.6. Proof of Theorem 3.6

Note that (11) still holds. With Lemma 6, we have

$$\mathbb{E}\|\mathbf{w}^K - \mathbf{w}^*\|^2 \leq C_4^2[1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2]^{2K}\|\mathbf{y}^1\|^2 + \gamma^2 C_4^2\sigma^2\sum_{i=1}^{K}[1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2]^{2K-2i}.$$

When $\Theta(\epsilon)$ and ϵ is small,

$$\gamma^2 C_4^2\sigma^2\sum_{i=1}^{K}[1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2]^{2K-2i} = C_4^2\sigma^2\frac{1-\beta}{\nu}\gamma + \mathcal{O}(\epsilon^2) = \mathcal{O}(\epsilon).$$

If $\mathbb{E}\|\mathbf{w}^K - \mathbf{w}^*\|^2 \leq \epsilon$, we then have

$$[1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2]^{2K} = \mathcal{O}(\epsilon).$$

The worst case is then $\mathcal{O}(\frac{\ln\frac{1}{\epsilon}}{\frac{\gamma\nu}{1-\beta} - C_3\epsilon^2}) = \widetilde{\mathcal{O}}(\frac{1-\beta}{\epsilon\nu})$.